

Міністерство освіти і науки України
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Кафедра комп'ютерних систем, мереж і кібербезпеки (№503)

ЗАТВЕРДЖУЮ

Керівник (гарант) освітньої програми


(підпис)

А. В. Шостак
(ініціали та прізвище)

« 30 » серпня 2024 р.

**РОБОЧА ПРОГРАМА ОБОВ'ЯЗКОВОЇ
НАВЧАЛЬНОЇ ДИСЦИПЛІНИ**

Технології великих даних
(назва навчальної дисципліни)

Галузь знань: 12 «Інформаційні технології»
(шифр і найменування галузі знань)

Спеціальність: 123 «Комп'ютерна інженерія»
(код та найменування спеціальності)

Освітня програма: «Системне програмування»
(найменування освітньої програми)

Форма навчання: денна

Рівень вищої освіти: перший (бакалаврський)

Харків 2024 рік

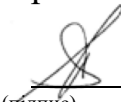
Робоча програма «Технології великих даних»
(назва дисципліни)

Розробник: Фесенко Г. В., професор, д.т.н., професор
(прізвище та ініціали, посада, науковий ступінь та вчене звання)  (підпис)

Робочу програму розглянуто на засіданні кафедри комп'ютерних систем, мереж і кібербезпеки
(назва кафедри)

Протокол № 1 від « 30 » 08 2024 р.

Завідувач кафедри д.т.н., професор
(науковий ступінь та вчене звання)


(підпис)

В. С. Харченко
(ініціали та прізвище)

1. Опис навчальної дисципліни

Найменування показника	Галузь знань, спеціальність, освітня програма, рівень вищої освіти	Характеристика навчальної дисципліни (денна форма навчання)
Кількість кредитів – 4	Галузь знань <u>12 «Інформаційні технології»</u> (шифр і найменування) Спеціальність <u>123 «Комп’ютерна інженерія»</u> (код і найменування) Освітня програма <u>«Системне програмування»</u> (найменування) Рівень вищої освіти: перший (бакалаврський)	Обов’язкова
Кількість модулів – 1		Навчальний рік
Кількість змістовних модулів – 2		2024/2025
Індивідуальне завдання <u>немає</u>		Семестр
Загальна кількість годин – 64/56		<u>7-й</u>
		Лекції*
Кількість тижневих годин для денної форми навчання: аудиторних – 4 самостійної роботи студента – 3,5		<u>32</u> години
		Практичні, семінарські*
		– годин
		Лабораторні*
	<u>32</u> годин	
	Самостійна робота	
	<u>56</u> години	
	Вид контролю	
	залік	

Співвідношення кількості годин аудиторних занять до самостійної роботи становить: для денної форми навчання – 64/56.

*Аудиторне навантаження може бути зменшене або збільшене на одну годину в залежності від розкладу занять.

2. Мета та завдання навчальної дисципліни

Мета вивчення: надання студентам знань та навичок щодо застосування технологій великих даних та розробки прикладного програмного забезпечення з їх використанням.

Завдання: придбання студентами знань про методи зберігання та обробки великих даних з використанням сучасних фреймворків та вмінь щодо їхнього застосування; придбання студентами знань про основні методи аналізу великих даних та технології використання нереляційних баз даних та вмінь щодо їхнього застосування; придбання студентами знань про технології паралельної та розподіленої обробки великих даних в пакетному та реальному режимах та вмінь щодо їхнього застосування.

Компетентності, які набуваються:

- (ЗК2) Здатність вчитися і оволодівати сучасними знаннями;
- (ЗК3) Здатність застосовувати знання у практичних ситуаціях;
- (ФК5) Здатність використовувати засоби і системи автоматизації проектування до розроблення компонентів комп'ютерних систем та мереж, Інтернет додатків, кіберфізичних систем тощо;
- (ФК9) Здатність системно адмініструвати, використовувати, адаптувати та експлуатувати наявні інформаційні технології та системи;

Очікувані результати навчання:

- (ПРН2) Мати навички проведення експериментів, збирання даних та моделювання в комп'ютерних системах;
- (ПРН3) Знати новітні технології в галузі комп'ютерної інженерії;
- (ПРН4) Знати та розуміти вплив технічних рішень в суспільному, економічному, соціальному і екологічному контексті;
- (ПРН6) Вміти застосовувати знання для ідентифікації, формулювання і розв'язування технічних задач спеціальності, використовуючи методи, що є найбільш придатними для досягнення поставлених цілей;
- (ПРН8) Вміти системно мислити та застосовувати творчі здібності до формування нових ідей;
- (ПРН11) Вміти здійснювати пошук інформації в різних джерелах для розв'язання задач комп'ютерної інженерії;
- (ПРН15) Вміти виконувати експериментальні дослідження за професійною тематикою;
- (ПРН19) Здатність адаптуватись до нових ситуацій, обґрунтовувати, приймати та реалізовувати у межах компетенції рішення;
- (ПРН20) Усвідомлювати необхідність навчання впродовж усього життя з метою поглиблення набутих та здобуття нових фахових знань, удосконалення креативного мислення.

Пререквізити – «Технології програмування», «Мобільне програмування», «Технології Java»

Кореквізити – «Тестування та забезпечення якості».

3. Програма навчальної дисципліни

Модуль 1.

Змістовий модуль 1. Основи великих даних

Тема 1. Розуміння великих даних

Термінологія та стандарти у галузі великих даних. Характеристики великих даних. Організаційні передумови, придбання, конфіденційність, безпека даних та походження даних. Особливі вимоги до управління, методологія, хмарні середовища. Життєвий цикл аналітики великих даних. Обробка транзакцій (OLTP) та аналітична обробка (OLAP) в реальному часі, добування, перетворення і завантаження (ETL). Сховища та вітрини даних.

Тема 2. Концепції та технології зберігання великих даних

Кластери, файлові системи та розподілені файлові системи, база даних NoSQL. Шардінг і реплікація. Поєднання шардінга і реплікації. Теорема CAP, принцип проектування бази даних ACID та принцип BASE. Дискові системи зберігання. Системи зберігання в оперативній пам'яті.

Тема 3. Екосистема Hadoop

Основні відомості про Hadoop та його екосистему. Взаємодія користувача з кластером Hadoop. Порівняння систем управління базами даних на основі SQL з Hadoop. Модель розподілених обчислень великих масивів даних MapReduce. Розподілена файлова система Hadoop. Менеджер ресурсів YARN.

Тема 4. Spark

Загальні відомості про Spark. Spark у процесі обробки та інженерії даних. Spark у наукових дослідженнях у галузі обробки даних. Фрейм даних. Ментальна модель Spark. Особливості використання фрейму даних. Взаємодія зі Spark. Стислий огляд компонент та взаємодій між ними. Порівняння Hadoop і Spark.

Модульний контроль

Змістовий модуль 2. Особливості побудови та застосування архітектур великих даних

Тема 5. Архітектури зберігання та обробки великих даних

Архітектура сховища великих даних тільки для читання: знайомство з архітектурою RDS; розподіл відповідальності команд та запитів; служби і компоненти сховища великих даних тільки для читання; рекомендації щодо використання RDS. Парадигма data lakehouse. Розроблення системного контексту для data lakehouse. Розроблення логічної архітектури data lakehouse. Розроблення принципів архітектури data lakehouse. Поглинання та обробка даних в data lakehouse. Зберігання та обслуговування даних у data lakehouse.

Забезпечення можливостей аналітики в data lakehouse. Впровадження управління даними в data lakehouse.

Тема 6. Функційна архітектура мережі, керованої великими даними

Загальні відомості про мережу, керовану великими даними. Функційна архітектура площини великих даних для мережі, керованої великими даними. Функційна архітектура площини мережі для мережі, керованої великими даними. Функційна архітектура площини управління для мережі, керованої великими даними. Приклади застосування мереж, керованих великими даними

Тема 7. Еталонна архітектура великих даних

Загальні відомості про еталонну архітектуру великих даних. Концепція еталонної архітектури великих даних. Представлення користувача. Наскрізні аспекти. Функційні компоненти еталонної архітектури великих даних.

Тема 8. Особливості корпоративних архітектур великих даних

Високорівнева та низькорівнева еталонні моделі архітектури великих даних. Архітектура великих даних від Microsoft. Особливості архітектур великих даних від University of Amsterdam, IBM та Oracle.

Модульний контроль

4. Структура навчальної дисципліни

Назва змістового модуля і тем	Кількість годин				
	Денна форма				
	Усього	У тому числі			
		л	п	лаб.	С.р.
1	2	3	4	5	6
Модуль 1					
Змістовий модуль 1					
Тема 1. Розуміння великих даних	11	4			7
Тема 2. Концепції та технології зберігання великих даних	17	4		6	7
Тема 3. Екосистема Hadoop: MapReduce, HDFS, YARN	17	4		6	7
Тема 4. Spark	14	3		4	7
Модульний контроль	1	1			
Разом за змістовим модулем 1	60	16		16	28
Змістовий модуль 2					
Тема 5. Архітектури зберігання та обробки великих даних	19	8		4	7
Тема 6. Функційна архітектура мережі, керованої великими даними	13	2		4	7
Тема 7. Еталонна архітектура великих даних	15	4		4	7
Тема 8. Особливості корпоративних архітектур великих даних	13	2		4	7

Модульний контроль	1	1			
Разом за змістовим модулем 2	60	16		16	28
Усього годин	120	32		32	56

5. Теми семінарських занять

№ п/п	Назва теми	Кількість годин
	<i>Не передбачено.</i>	
	Разом	

6. Теми практичних занять

№ п/п	Назва теми	Кількість годин
1	<i>Не передбачено.</i>	
	Разом	

7. Теми лабораторних занять

№ п/п	Назва теми	Кількість годин
1	Програмування операцій зі стійкими розподіленими наборами даних (Resilient Distributed Datasets, RDD) з використанням Spark	6
2	Створення та робота з DataFrames з використанням Spark	6
3	Робота з графами з використанням Spark GraphX	4
4	Реалізація алгоритмів машинного навчання з використанням Spark MLlib і пакета spark.ml: статистичні обчислення і перевірка гіпотез	4
5	Реалізація алгоритмів машинного навчання з використанням Spark MLlib і пакета spark.ml: вилучення (TF-IDF, Word2Vec, CountVectorizer і CountVectorizerModel) і перетворення (Tokenizer, StopWordsRemover, n-gram, Binarizer, StringIndexer, IndexToString) ознак	4
6	Реалізація алгоритмів машинного навчання з використанням Spark MLlib і пакета spark.ml: перетворення (Interaction, MinMaxScaler, MaxAbsScaler, Bucketizer, ElementwiseProduct, SQLTransformer, VectorAssembler, Imputer) і вибір (VectorSlicer, MinHash) ознак	4
7	Реалізація алгоритмів машинного навчання з використанням Spark MLlib і пакета spark.ml: кластеризація і колаборативна фільтрація	4
	Разом	32

8. Самостійна робота

№ п/п	Назва теми	Кількість годин
1	Загальні принципи аналізу даних за допомогою Spark	7
2	Особливості роботи зі стійкими розподіленими наборами даних (RDD)	7
3	Особливості роботи с парами «ключ-значення»	7
4	Пакет для роботи зі структурованими даними Spark SQL	7
5	Бібліотека для обробки графів та виконання паралельних обчислень Spark GraphX	7
6	Компонент для роботи з потоковими даними Spark Structured Streaming	7
7	Бібліотека Spark Mlib для реалізації машинного навчання	7
8	Підсистема безпеки великих даних еталонної архітектури великих даних від NIST	7
	Разом	56

9. Індивідуальні завдання

Індивідуальні завдання не передбачені.

10. Методи навчання

Проведення аудиторних лекцій, лабораторних занять, консультацій, а також самостійна робота здобувачів освіти за матеріалами, опублікованими кафедрою.

11. Методи контролю

Проведення поточного тестового контролю, тестового модульного контролю, підсумковий контроль у вигляді заліку.

12. Критерії оцінювання та розподіл балів, які отримують студенти

12.1. Розподіл балів, які отримують студенти (кількісні критерії оцінювання)

Складові навчальної роботи	Бали за одне заняття (зав- дання)	Кіль- кість занять (завдань)	Сумарна кіль- кість балів
Змістовий модуль 1			
Робота на лекціях	0...1	8	0...8
Лабораторні заняття	0...6	3	0...18

Модульний контроль	0...20	1	0...20
Змістовий модуль 2			
Робота на лекціях	0...1	8	0...8
Лабораторні заняття	0...6	4	0...24
Модульний контроль	0...20	1	0...22
Усього за семестр			0...100

Білет для заліку складається з двох теоретичних і одного практичного запитання. За теоретичні запитання студент отримує до 60 балів (до 30 балів за кожне) , за практичне – до 40 балів.

Під час складання семестрового іспиту здобувач має можливість отримати максимум 100 балів.

Критерії оцінювання роботи студента протягом семестру

Задовільно (60-74). Показати мінімум знань та умінь. Захистити не менше 80% від усіх завдань лабораторних занять. Вміти самостійно давати характеристику використовуваній технології зберігання, обробки та аналізу великих даних, вміти програмування операції зі стійкими розподіленими наборами даних та DataFrames з використанням Spark.

Добре (75-89). Твердо знати мінімум знань, виконати не менше 90% завдань лабораторних занять. Вміти пояснювати вибір технологій зберігання, обробки та аналізу великих даних, вміти програмувати операції зі стійкими розподіленими наборами даних, DataFrames, обробляти графи з використанням Spark. Розуміти корпоративні технології і бізнес-аналітику для великих даних та вміти застосовувати їх для вирішення практичних завдань.

Відмінно (90 - 100). Повно знати основний та додатковий матеріал. Виконати усі завдання лабораторних занять в обумовлені викладачем строки. Знати усі теми. Орієнтуватися у підручниках та посібниках. Досконально знати усі основні технології та фреймворки, які використовуються для зберігання, обробки та аналізу великих даних. Вміти удосконалювати корпоративну технологію та бізнес аналітику для великих даних. Вміти програмувати операції зі стійкими розподіленими наборами даних, DataFrames, обробляти графи з використанням Spark, реалізовувати алгоритми машинного навчання з використанням бібліотеки Spark MLlib.

Шкала оцінювання: бальна і традиційна

Сума балів	Оцінка за традиційною шкалою	
	Іспит, диференційований залік	Залік
90 – 100	Відмінно	Зараховано
75 – 89	Добре	
60 – 74	Задовільно	
0 – 59	Незадовільно	Не зараховано

13. Методичне забезпечення

1. Сторінка дисципліни у системі дистанційного навчання «Ментор» [Ел. ресурс]. URL: <https://mentor.khai.edu/course/view.php?id=1600>.

14. Рекомендована література

Базова

1. Олещенко Л. М. *Технології оброблення великих даних. Конспект лекцій*. Київ : КПІ ім. Ігоря Сікорського, 2021. 227 с.
2. Balusamy B., Abirami N., Kadry S., Gandomi A. *Big Data: Concepts, Technology, and Architecture*. USA, NJ, Hoboken : Wiley, 2021. 371 p.
3. Pop F., Neagu G. *Big Data Platforms and Applications: Case Studies, Methods, Techniques, and Performance Evaluation*. Switzerland, Cham : Springer, 2021. 300 p.
4. Perrin J.-G. *Spark in action*. USA, NY, Shelter Island : Manning Publications Co., 2020. 576 p.

Допоміжна

1. Ravi V., Cherukuri A. *Handbook of Big Data Analytics*. Vol. 1. Methodology. UK, London : The Institution of Engineering and Technology, 2021. 390 p.
2. Ravi V., Cherukuri A. *Handbook of Big Data Analytics*. Vol. 2. Applications in ICT, security and business analytics. UK, London : The Institution of Engineering and Technology, 2021. 420 p.
3. Chinnici M., Pop F., Neagu G. *Data Science and Big Data Analytics in Smart Environments*. USA, Boca Raton, FL : CRC Press, 2021. 305 p.

15. Інформаційні ресурси

1. The 9 Best Free Online Big Data and Data Science Courses [Ел. ресурс]. URL: <https://bernardmarr.com/the-9-best-free-online-big-data-and-data-science-courses>.
2. Big Data and Data Science Courses [Ел. ресурс]. URL: <https://www.edx.org/learn/big-data>.
3. Mastering Big Data Analytics [Ел. ресурс]. URL: <https://www.mygreatlearning.com/academy/learn-for-free/courses/mastering-big-data-analytics>.